CITEFUSION-RAG: HYBRID FUSED RETRIEVAL WITH CITATION-FIRST PROMPTING FOR ACCURATE DOCUMENTATION QA

Anonymous authors
Paper under double-blind review

ABSTRACT

We present CiteFusion-RAG, a retrieval-augmented generation (RAG) pipeline for question answering over technical documentation that combines hybrid retrieval with citation-first prompting. The retrieval stage fuses sparse BM25 scores and dense embedding scores with lightweight entity hints, while overlapping chunks and neighbor expansion preserve evidence across section boundaries. The generation stage instructs the language model to ground claims in labeled snippets, cite supporting sources, and abstain when the retrieved evidence is insufficient. We evaluate the pipeline on five documentation corpora (Langfuse, Mastra, Langflow, Bland, and Firecrawl), each containing 150 question-answer pairs balanced across easy, medium, and hard difficulty levels. Using answer relevancy as the primary automatic metric, CiteFusion-RAG achieves a reported macro-average of 88.43%, approximately 11 percentage points above the strongest evaluated baseline. On hard questions, it achieves 84.35%, approximately 12.1 points above that baseline. These results suggest that combining complementary retrieval signals with explicit attribution instructions can improve answer relevancy in technical-documentation QA. Broader faithfulness evaluation and human assessment remain necessary to establish claim-level correctness. Evaluation scripts, the benchmark, and a simplified reference implementation are planned for release upon camera-ready publication.

AI USE STATEMENT

The evaluated pipeline uses `o3-mini` for answer generation. A generative AI assistant was used to draft and revise the wording and format of this abstract from the existing manuscript. The reported experimental results originate from that manuscript; no new experiments were conducted in preparing this abstract.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.